Automatic Identification of Support Verbs: A Step Towards a Definition of Semantic Weight

Mark Dras *
65 Epping Road, North Ryde NSW 2113, Australia
Email: t-markdr@microsoft.com

Abstract

Current measures of the readability of texts are very simplistic, typically based on counts of words or syllables per sentence. A more sophisticated analysis needs to take account of the fact that the particular distributions of meanings across wordings chosen by the writer, and the consequent variations in syntactic structure, have a significant effect on readability.

A step towards the required sophistication is provided by the notion of LEXICAL DENSITY (Halliday, 1985), which suggests that different words carry different amounts of semantic weight; this idea of semantic weight is also used implicitly in areas such as information retrieval and authorship attribution.

Current definitions of these notions of lexical density and semantic weight are based on the division of words into closed and open classes, and on intuition. This paper develops a computationally tractable definition of semantic weight, concentrating on what it means for a word to be semantically light; the definition involves looking at the frequency of a word in particular syntactic constructions which are indicative of lightness. Verbs such as *make* and *take*, when they function as support verbs, are often considered to be semantically light. To test our definition, we carried out an experiment based on that of Grefenstette and Teufel (1995), where we automatically identify light instances of these words in a corpus; this was done by incorporating our frequency-related definition of semantic weight into a statistical approach similar to that of Grefenstette and Teufel. The results show that this is a plausible definition of semantic lightness for verbs, which can possibly be extended to defining semantic lightness for other classes of words.

*Reprinted with kind permission from: "Automatic Identification of Support Verbs: A Step Towards a Definition of Semantic Weight" in Proceedings of the Eighth Australian Joint Conference on Artificial Intelligence (World Scientific, Singapore, 1995) pp 451 - 458.

1 Introduction

There are a number of ways of measuring properties of text, and from there proceeding to make stylistic judgments; they can be found in style guides, and include calculating readability indices, counting the number of passive constructions, and so on. One attribute of text that is rarely mentioned explicitly in these style guides, but which underpins many of the pieces of advice, is that of SEMANTIC DENSITY (see Dras and Dale, 1995). Consider the following pair of sentences, taken from Halliday (1985):

(1) a. Slavish imitation of models is nowhere implied.

b. It is not implied anywhere that there are models which should be slavishly imitated.

It is apparent that the first of the pair is 'denser' than the second: both express the same (propositional) meaning, but the first does so in a more compact way. Halliday terms this LEXICAL DENSITY, 'the density with which information is presented' (p68), and measures it by looking at the proportion of CONTENT WORDS. Halliday adopts a fairly standard conception of content words as those which belong to the open word classes: nouns, verbs, adjectives and so on. Non-content words are then those that belong to the closed classes, such as prepositions, auxiliaries and so on. The non-content, closed class words are viewed by Halliday as lacking in informational content.

He does, however, note that there are some words on the borderline between content and non-content words which are lexical items but in many cases do little more than perform a grammatical function. These include the noun *thing*, as in *That's a thing I could do without* (which could be rewritten as *I could do without that*), and the verb *make*, as in *Christophe made a decision to come to the Drag Day* (possibly rewritten as *Christophe decided to come to the Drag Day*).

In this paper I look at a possible definitional extension of non-content words, which incorporates the intuition expressed by Halliday that words like *make* often contribute little, if any, propositional meaning to the text. This new definition is tested by an experiment modelled on that of Grefenstette and Teufel (1995), which tries to find the support verb that particular nominalisations will take—why *decision*, for example, takes *make*, and not *have*, *do*, *eat* or *perambulate*.

2 Lightness of Words

2.1 Current views of lightness

The fact that a word contributes little if any content to a text is used in a number of areas of linguistics and computational linguistics. In information retrieval, non-content words are discarded, as they cannot help to identify the topic of a text. Mosteller and Wallace (1984),

on the other hand, retain them and discard the content words when attempting to statistically determine the authorship of the disputed Federalist papers, reasoning that while content words may vary across topic, for a given author non-content words will not. Halliday (1985), as mentioned above, uses them to define the informational density of a text, in order to compare spoken and written text.

What comprises the class of non-content words is neither uniform nor clearly defined. Halliday defines it to be the set of those words which are part of a closed class system; information retrieval commonly uses a combination of high-frequency and known function words; Mosteller and Wallace use a list which was derived from sources such as the King James Bible.

Halliday proposes that relative frequency of a word can be used to indicate the amount of information it contributes. If this is true, the choice of closed class words to represent non-content words is plausible, since a given grammatical item (*the*, *and*, *it*) is more likely to have a higher frequency of occurrence than a given lexical item (*dog*, *run*, *verisimilitude*). It would also include *make* and *thing*, which are high frequency lexical items.

However, this idea needs to be further refined. A quick inspection of a corpus will show that there are a number of words with definite propositional content which rank above non-content words in frequency. In the 8 million word Grolier's Encyclopedia the verb *include* (which definitely conveys information, so it can't be a light constituent) occurs transitively 4284 times, as against *make*'s 2697 times.

2.2 A different view

Make does, however, occur more frequently in constructions which I will call LIGHT CONSTRUCTIONS, such as *make a decision*; they are mentioned under one name or another by linguists and style guide authors, and what characterises them is that the light constituent can be deleted (with some rewriting of the remaining text to retain grammaticality). For example, *make a decision* can be rewritten as *decide*, the light element being *make*. Jespersen (1954) is one of the earliest to note these, commenting on the LIGHT VERBS in expressions such as *take a walk*. Style guide writers like Kane (1983) mention 'deadwood' which can be eliminated from phrases such as *It is important for teachers to have a knowledge of their students* (a possible rewriting being *It is important for teachers to know their students*). There are quite a few of these constructions, such as light verbs with noun phrase complements, light verbs with adjectival complements, and light nouns with post-modifiers (Dras and Dale, 1995), but this paper only looks at one construction, the light verbs with NP complements. By definition, these constructions will characteristically contain light verbs; this paper therefore proposes that a modified definition be used for indicating whether a word can be considered a non-content one: that the word has a high relative frequency in these light constructions. In particular, it examines the relationship of the relative frequency of a verb in these light verb constructions to its lightness.

It has been suggested that semantic factors are what determine the relationship between

a syntactic construction and its associated light verb. Wierzbicka (1982) proposes a set of semantic rules for determining the light verb that corresponds to a particular noun object—an explanation of why one can *have a drink* but not **have an eat*. However, defining these rules by hand for all nouns would be too time-consuming to be practical. Grefenstette and Teufel (1995) take a statistical approach to finding what is termed the SUPPORT VERB for a particular noun. They look at several nouns, including *appeal*, *proposal*, and *demand*. In a corpus of newspaper articles, they look for occurrences of the noun and corresponding verb to find the most likely candidate for the support verb. They find that the most likely support verb for *appeal* is *make*, which accords with intuition, but for *proposal*, their system also finds *reject* as an equally likely candidate; and for *demand*, the most likely candidate is *meet*. In this paper, I conduct a similar experiment, finding support verbs for given nouns, to test the definition proposed above: that a word's status regarding content-freeness is related to its frequency of occurrence in light constructions.

3 Experiment

The aim of the experiment is to show that there is a relationship between relative frequency of verbs in particular constructions and the content-freeness of these verbs. A consequence of this is to be able to choose the light verb that corresponds to a nominalisation in a light verb–NP complement construction—the nominalisation's support verb (SV). Deverbal nominalisations are chosen as they are the kinds of grammatical entities which enter into the SV–NP complement construction.

3.1 Experiment design

A way of extracting light verbs from a corpus is to simply take all verb-object pairs where the object is a deverbal nominalisation. Grefenstette and Teufel use only LOCAL INFORMATION, information that is specific to a particular nominal. Counting all occurrences of each noun in verb-object pairs yields a local relative frequency for each verb with respect to that noun. So, to determine the support verb for *proposal* they look only at verbs which co-occur with the noun *proposal*. While it seems intuitively obvious to native English speakers that *make* is a more likely candidate for support verb than *reject*, the local frequency evidence does not indicate this. Speakers also use the fact that *make* is the support verb for other nominalisations such as *judgment* and *decision*. I have termed this knowledge GLOBAL INFORMATION. Counting all occurrences of each verb, regardless of their objects, yields a global relative frequency for that verb. In this experiment the local information is combined with the global information to produce a modified likelihood of being a support verb.

3.2 Deriving local and global information

To gather local and global information, the 1992 version of Grolier's encyclopedia, tagged by the part-of-speech tagger developed by Brill (1993), was used. A heuristic for producing the local information involved searching the corpus for the nominal, determining the verb (if any) for which the nominal was the direct object, and measuring the relative frequency of these verbs.

The theoretical global information is a measure of how productive a given support verb is: that is, how many different instances of the SV-NP construction it enters into. The more productive verbs (like *make*) rank higher on the list than less productive verbs (like *bear*); this is combined with the local information so that the more productive verbs, for a particular nominal, are subsequently ranked more highly than by the local information alone. This weighting technique is similar to that used by Yarowsky (1992) in the context of sense disambiguation. In his work he uses counts of words in a window around a key word to determine the salience of this key word to a particular sense. These word counts are weighted so that more common words contribute less; that is, the less common words are accorded more importance. We, on the other hand, want to give more importance to the more common words, given our assumption that it is high relative frequency in particular constructions that helps define semantic lightness.

Grefenstette and Teufel note that a confounding factor in the local information, when picking out nominals and their associated verbs, is that the nominal may have become CONCRETISED. Generally, nominals represent an abstract concept, being essentially events represented in noun form; but it is possible for the nominal to represent a physical embodiment of that concept. For example:

(2) a. He made his formal proposal to the full committee.

b. He put the proposal in the drawer.

The abstract and concretised versions will tend to have different associated verbs. However, if the relative frequency hypothesis is true, and the global information is an accurate reflection of the innate lightness of a verb, this will elevate the light verb over the 'heavier' ones, which will be more likely to be associated with the concretised forms.

In practice, the global information is calculated from the aggregate of the local data. This means that there is a lot of noise—all of the incorrect candidates for support verb are included in the global information—but again, if the relative frequency hypothesis is true, the relative frequency of the support verb in the local information will be high (although not necessarily the highest), while this is not true for non-support verbs. So aggregating all of these should reinforce the support verbs and not the others.

3.3 Generating nominalisations

To construct the global information, a comprehensive list of deverbal nominalisations is needed, together with the associated support verbs, determined from the local information. To generate this list in a partially automated manner, Longman's Dictionary of Contemporary English (LDOCE) was used, including both built-in information and a heuristic: a nominal is an event represented in noun form, so the procedure used here for deriving a list of them involved looking for nouns with associated STEM VERBS.

Some verbs have this information encoded in their entries: for example, *adjust* lists *adjustment* as its nominalisation; there were 257 verbs in this category. For others, an automatic orthographic heuristic that matched nouns with verbs was manually filtered to produce 1414 more deverbal nominalisations. A system to identify support verbs for nominalisations was implemented by tabulating all the verbs for which these nominals were the direct object.

The list of nominals did not cover some of the nominals from the test set (listed in the table below). The local information was generated for each of the excluded test set nominals and aggregated into the global information. Candidates for support verb were ranked in order of the product of local and global relative frequency of each candidate verb.

3.4 The test set and results

The light verb constructions and their constituent nominals used for testing were taken from a range of sources, so that they would not be biased to one particular genre.

The test set and results are summarised in Table 1; the table contains:

- the source text;
- the corresponding verb, which the source can be rewritten as;
- the reference for the source text;
- the system's first choice candidate for support verb for the source text's constituent nominalisation;
- the system's second choice; and
- the ratio of the adjusted frequency, which is defined as the product of local and global relative frequencies, for the first and second choices.

3.5 Discussion

Of the 18 examples, 13 of the choices for support verb match the corresponding one from the source text. Of the five incorrect ones, three were incorrect because of lack of data:

Source Text	Verb	Reference	Choice 1	Choice 2	Ratio
make an attempt	attempt	[2]	make	include	9.36
make a change	change	[2]	make	produce	1.85
make a concession	concede	[2]	make	include	11.47
make a demand	demand	[3]	make	create	1.03
make a distinction	distinguish	[2]	make	have	3.04
have a drink (of)	drink	[10]	become	N/A	N/A
have an effect (on)	affect	[2]	have	produce	3.04
have a feeling	feel	[5]	have	produce	3.27
make a gift (of)	give	[5]	have	include	9.89
do harm (to)	harm	[6]	cause	do	1.26
make a judgment	judge	[2]	make	have	2.43
have a knowledge (of)	know	[8]	have	use	12.36
make progress	progress	[5]	make	allow	64.33
make a proposal	propose	[3]	make	include	1.10
bear a resemblance (to)	resemble	[6]	bear	have	2.64
give a shove (to)	shove	[5]	N/A	N/A	N/A
have a snooze	snooze	[5]	N/A	N/A	N/A
make use (of)	use	[2]	make	have	6.55

Table 1: Support verb candidates

there are no occurrences of *snooze* or *shove* as direct objects of verbs in Grolier's, most probably because they belong to a more informal register than that used in encyclopedias. Similarly, *have a drink* is an informal phrase that would not normally be found in an encyclopedia.

Another of the incorrect cases, *harm*, had *cause* as the proposed alternative. This is an equally valid support verb, and in any case, *do* was the second choice by only a small margin. This is true for a number of cases: where there is an alternative support verb to the one used in the source text, the second alternative represents another plausible choice, and the frequency ratio margin is small (for example, for *change* and *resemblance*).

So if only the cases where enough data exists are considered, and if alternative support verbs are allowed, the success rate becomes 14 of 15.

4 Conclusion

It is apparent that what constitutes a valid light verb construction depends on the genre and register of the text. More accurate results could no doubt be obtained by using a corpus that was more representative of general English. Also, where shortcuts were taken (for example, by not removing concretised nominalisations), more precision could be

obtained. Notwithstanding these considerations, using the relative frequency of a verb in light constructions seems to be a fairly good indicator of a verb's content-freeness, providing plausible choices for support verbs for nominalisations.

Further work will involve extending this definition to other light constructions—light verbs with adjectival complements, and light nouns with post-modifiers—and fitting closed-class words, which traditionally comprise the class of non-content words, into the framework. This framework can then be used to give a more accurate indication of lexical density; to more accurately choose words to leave out, or include in, in the fields of information retrieval and authorship attribution; and to fine-tune stylistic judgments.

5 Acknowledgements

This work has benefited from the kind assistance of Robert Dale and Mark Lauer. I'd also like to thank Mike Johnson for many fruitful discussions. Financial support is gratefully acknowledged from the Australian Government and the Microsoft Institute.

References

[1] Brill, Eric. 1993. *A Corpus-based Approach to Language Learning*. PhD Thesis, University of Pennsylvania, PA.

[2] Dras, Mark and Robert Dale. 1995. *Style and Semantic Density*. Microsoft Institute Research Report no. 95-04.

[3] Grefenstette, Greg and Simone Teufel. 1995. "Corpus-based Method for Automatic Identification of Support Verbs for Nominalizations". To appear in Proc. of EACL'95.

[4] Halliday, Michael A. K. 1985. *Spoken and Written Language*. Oxford University Press, Oxford.

[5] Harris, Zellig. 1957. "Co-occurrence and transformation in linguistic structure". *Language*, 33, 293-340.

[6] Huddleston, Rodney. 1968. *Sentence and Clause in Scientific English*. Communication Research Centre, University College, London.

[7] Jespersen, Otto. 1942. *A Modern English Grammar on Historical Principles, Vol VI*. Allen and Unwin, London.

[8] Kane, Thomas S. 1983. *The Oxford Guide to Writing*. Oxford University Press. New York, NY.

[9] Mosteller, Frederick and David Wallace. 1984. *Applied Bayesian and Classical Inference: the Case of the Federalist Papers*. Springer-Verlag. New York, NY.

[10] Wierzbicka, Anna. 1982. "Why Can You *Have a Drink* When You Can't **Have an Eat*?". *Language*, 58(4), 753-799.

[11] Yarowsky, D. 1992. "Word sense disambiguation using statistical models of Roget's categories trained on large corpora". *Proceedings of the 14th International Conference on Computational Linguistics*, 454-460.